

A Knowledge-Oriented Framework for Gene Selection

Zheng Zhao[†], Jiangxin Wang[‡], Shashvata Sharma[†], Nitin Agarwal[†], Huan Liu[†], Yung Chang[‡]

[†]Computer Science and Engineering Department, Arizona State University; [‡]School of Life Science, CIDV, The Biodesign Institute, Arizona State University

{zhaozheng, jiangxin.wang, shashvata.sharma, nagarwa6, huan.liu, yung.chang}@asu.edu



1. Motivation

Traditional gene selection based on cDNA microarray suffers several problems.

- small sample size;
- cannot distinguish “trigger” from “fire”;
- biological relevance \neq statistical relevance;

Figure 1: Gene selection on ALL data

Unsupervised (Accuracy: 0.61)			
SFRS5	TM9SF1	WTAP	GPSM3
STAC3	POMP	SLC25A6	
Supervised (Accuracy: 0.97)			
USP33	IL2RG	SIGIRR	CHCHD2

Recent development in bioinformatics has made various knowledge sources available such as:

- mRNA expression profiles;
- gene sequence & gene annotation;
- biological pathway;
- literature, etc.

Efficient and effective integration of knowledge from multiple sources for gene selection helps on solving these problems and enables to select genes bearing significant biological relevance.

2. Contribution

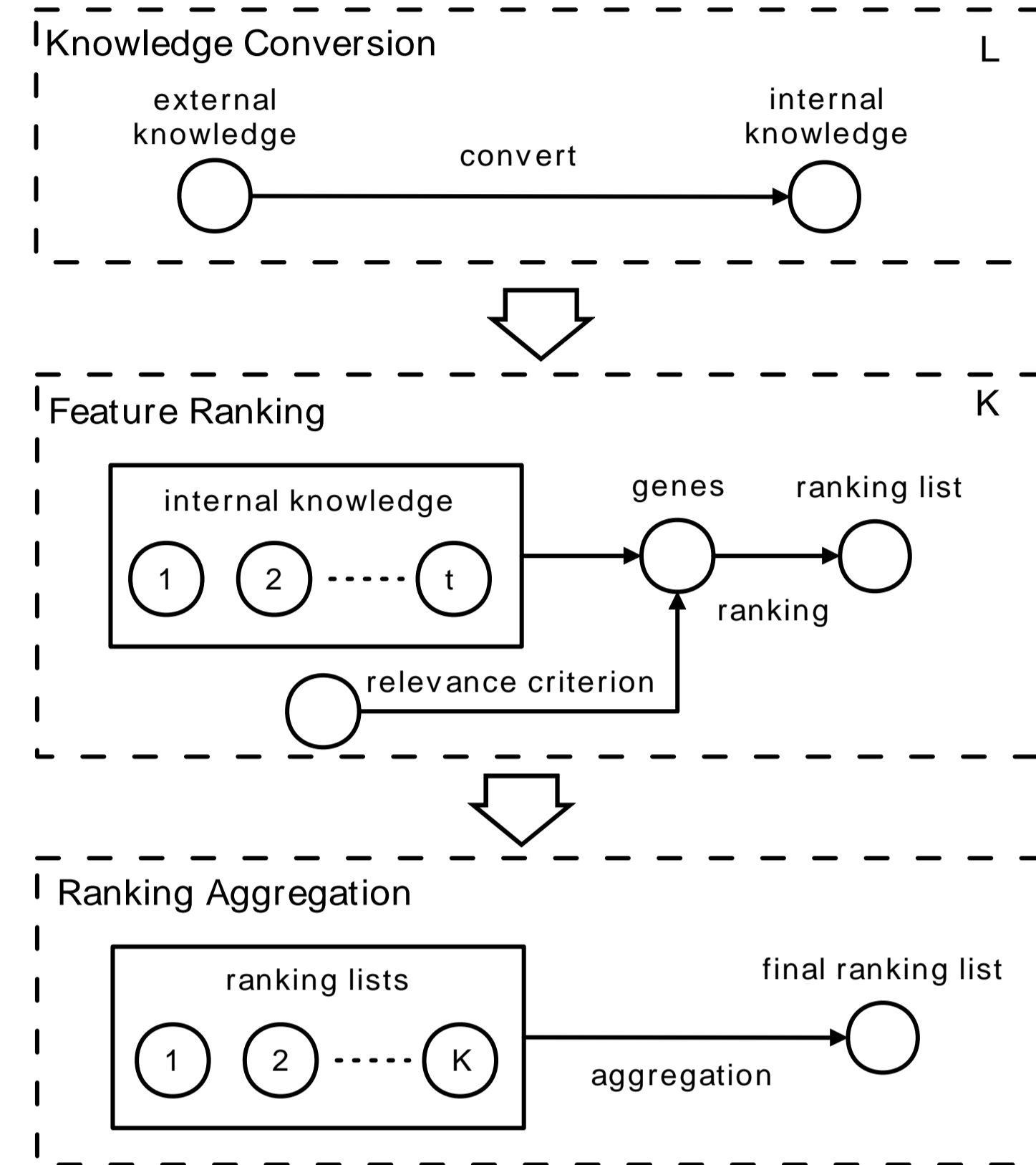
We proposed a novel framework for integrating different types of knowledge for gene selection.

- we categorized different types knowledge sources;
- we proposed a two layer (external/internal) framework for knowledge representation and management;
- we studied how ranking genes with different types of internal knowledge and ranking aggregation for knowledge integration;
- we conducted extensive experiments for evaluation;

3. A Framework for Knowledge-Oriented Gene Selection

We developed a general framework for systematically integrating different types of knowledge to achieve Knowledge-Oriented Gene Selection, thus it is named KOGS. It contains three major steps: (1) **Knowledge Conversion**, $\mathcal{K}_i^{int} = c_i(\mathcal{K}_i^{ext})$: converting external knowledge (from human) to internal knowledge that can be used by gene selection algorithms. Here, we have L different external knowledge sources $\mathcal{K}_1^{ext}, \dots, \mathcal{K}_L^{ext}$. And for the i th external knowledge, an operator $c_i(\cdot)$ can be applied to convert \mathcal{K}_i^{ext} to the internal knowledge \mathcal{K}_i^{int} . (2) **Feature Ranking**, $R_i^{rank} = \mathcal{R}(KNOW_i, C_i, G)$: here we use K sets of internal knowledge $KNOW_1, \dots, KNOW_K$ to rank genes, where $KNOW_i$ is defined as: $KNOW_i = \{\mathcal{K}_{i_1}^{int} \dots \mathcal{K}_{i_{t_i}}^{int}\}$. Let C_i be a relevance criterion, $G = \{g_1, \dots, g_M\}$ be a set of M genes and $\mathcal{R}_i(\cdot)$ be a gene ranking function. (3) **Ranking Aggregation**, $R_F^{rank} = \mathcal{A}(R_1^{rank}, \dots, R_K^{rank}, C)$: after obtaining the K ranking lists, they need to be integrated to obtain a final ranking to estimate the relevance of the genes. Here $\mathcal{A}(\cdot)$ is an aggregating operator and C be an aggregation criterion.

Figure 2: The framework of KOGS.



EXTERNAL KNOWLEDGE

Various external knowledge sources can be categorized into two groups: the knowledge about genes, and the knowledge about samples, which can be further categorized into five categories:

- Knowledge of gene similarity, \mathcal{K}_{SIM}^{ext} ;
- knowledge of gene functions, \mathcal{K}_{FUN}^{ext} ;
- knowledge of gene interaction, \mathcal{K}_{INT}^{ext} ;
- sample categories, \mathcal{K}_{CAT}^{ext} ;
- samples' geometric relationship, \mathcal{K}_{GEO}^{ext} .

Category	Knowledge	Class Label, Sample Hierarchy
Samples	\mathcal{K}_{CAT}^{ext} - Category	Class Label, Sample Hierarchy
	\mathcal{K}_{GEO}^{ext} - Geometry	mRNA Expression Profile, mRNA Expression Profile
Genes	\mathcal{K}_{SIM}^{ext} - Similarity	Gene Sequence, Gene Ontology Annotation, Gene Lineage
	\mathcal{K}_{FUN}^{ext} - Function	Gene Ontology, Metabolic Pathway, Gene-Disease Association
	\mathcal{K}_{INT}^{ext} - Interaction	Metabolic Pathway, Protein-Protein Interaction

INTERNAL KNOWLEDGE

Based on a), whether the definition facilitates certain types of external knowledge to be easily converted to its form. b), whether it can be effectively used to rank genes, in KOGS, we use the following types of knowledge:

- sample category, \mathcal{K}_{CAT}^{int} ;
- sample geometric pattern, \mathcal{K}_{GEO}^{int} ;
- gene connection, \mathcal{K}_{CON}^{int} ;
- gene function, \mathcal{K}_{FUN}^{int} .

KNOWLEDGE CONVERSION

External Knowledge	Internal Knowledge
$\mathcal{K}_{GEO}^{ext}, \mathcal{K}_{FUN}^{ext}, \mathcal{K}_{SIM}^{ext}$	\mathcal{K}_{GEO}^{int}
$\mathcal{K}_{SIM}^{ext}, \mathcal{K}_{INT}^{ext}$	\mathcal{K}_{CON}^{int}
\mathcal{K}_{FUN}^{ext}	\mathcal{K}_{FUN}^{int}
\mathcal{K}_{CAT}^{ext}	\mathcal{K}_{CAT}^{int}

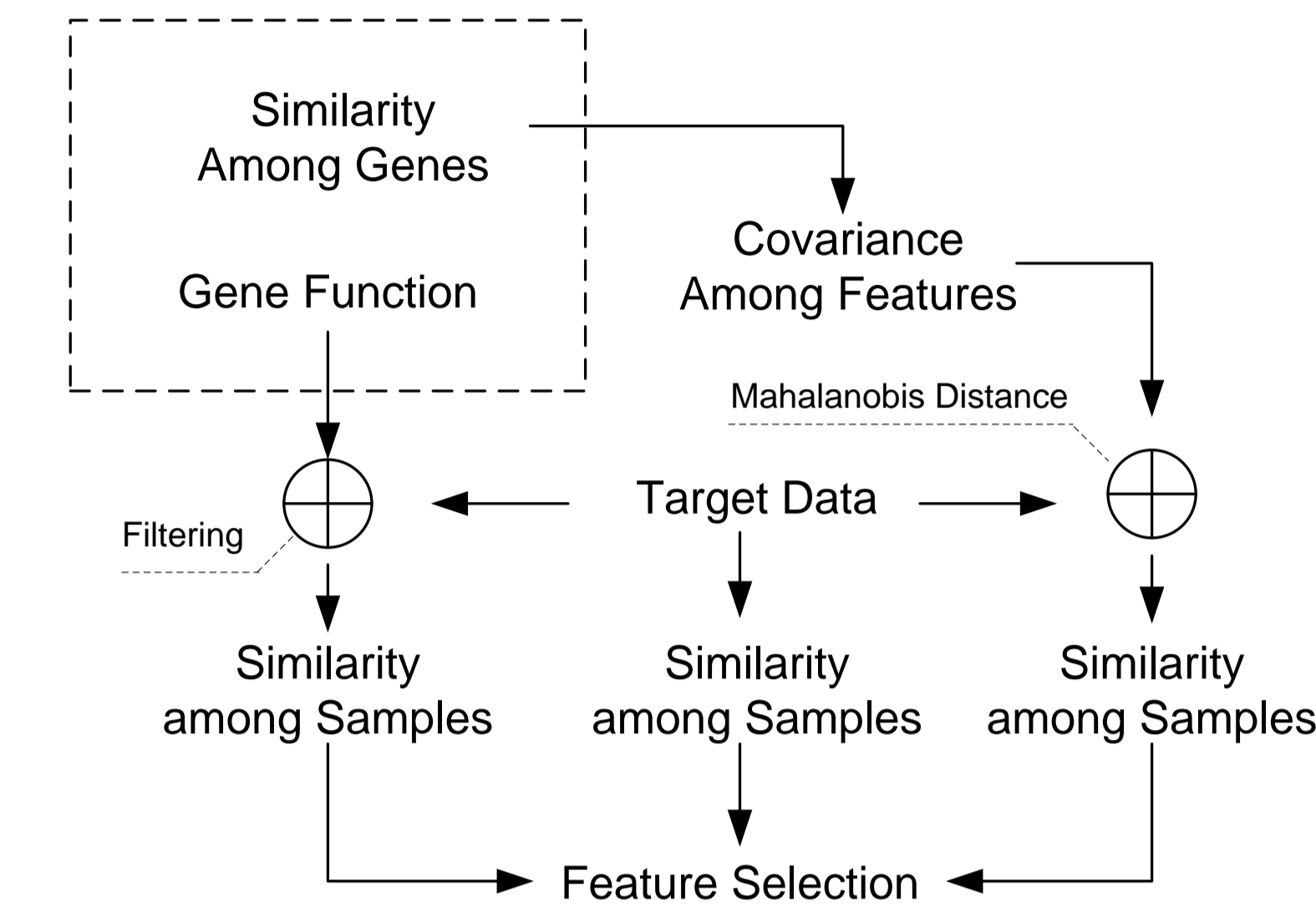


Figure 5: An example of knowledge conversion

Details of knowledge conversion can be found in the report.

4. Ranking Genes with Knowledge

Different types of internal knowledge can be used to rank genes and the obtained ranking lists can be combined to obtain a final list.

RANKING USING INTERNAL KNOWLEDGE

- \mathcal{K}_{CAT}^{int} + Traditional Supervised Gene Selection.
- \mathcal{K}_{CON}^{int} + Relevance Propagation: $\mathbf{r}^* = (I - \lambda P)^{-1} \mathbf{r}_{int}$.
- \mathcal{K}_{FUN}^{int} + Functional Relevance Voting: $r_i = \mathbf{f}_i^T \mathbf{r}_{fun}$.
- \mathcal{K}_{GEO}^{int} + Geometric Consistency: $r_i = \left(\frac{\hat{\mathbf{g}}_i^T \gamma(\mathcal{L}) \hat{\mathbf{g}}_i}{1 - (\hat{\mathbf{g}}_i^T \xi_0)} \right)^{-1}$.

AGGREGATING GENE RANKING LISTS

We proposed a probabilistic model for ranking aggregation. Let g_i denote gene i , and its rank in ranking list l be $r_{l,i}$, the probability of g_i to be relevant according to ranking list l is defined as:

$$P(r_{l,i}) = \frac{1}{B} \exp\left(\frac{1}{r_{l,i}}\right), B = \sum_{j=1}^M \exp\left(\frac{1}{j}\right). \quad (1)$$

Let $\pi_l = P(R_l)$ be the probability that ranking list l is used to rank genes, the probability of g_i to be relevant can be calculated by marginalizing the joint probability $P(g_i, R_l)$.

$$P(g_i) = \sum_{l=1}^L P(r_{l,i}) P(R_l) = \sum_{l=1}^L P(r_{l,i}) \pi_l. \quad (2)$$

The π_l can be computed effectively via an EM algorithm.

5. Experimental Study

To evaluate the performance of KOGS, we test gene selection using various single knowledge source as well as using multiple knowledge sources together on the ALL data*. The data contains the expression profiling of 4,670 genes in bone marrow from pediatric 18 patients with acute lymphoblastic leukemia (ALL). The data provides insight into the pathogenesis of childhood ALL. Five different knowledge sources are used in the experiments: (1) Sample Category, (2) Gene Expression, (3) Metabolic Pathway,

(4) Cancer-Gene Annotation, (5) Gene Ontology (GO) Annotation. The details of the experiment setup can be found in the following two tables, respectively.

METHODS	KNOWLEDGE SOURCES	EXT. KNW. CAT.	INT. KNW. CAT.	RANKING CRITERION
SPEC	Expression	\mathcal{K}_{GEO}^{ext}	\mathcal{K}_{GEO}^{int}	Geometric Consistency
Fisher Score	Category	\mathcal{K}_{CAT}^{ext}	\mathcal{K}_{CAT}^{int}	Supervised Gene Selection
Pathway-FILT	Pathway	\mathcal{K}_{FUN}^{ext}	\mathcal{K}_{FUN}^{int}	Geometric Consistency
GO-REL-VOTE	GO Anno	\mathcal{K}_{FUN}^{ext}	\mathcal{K}_{FUN}^{int}	Functional Relevance Voting
GO-MAH	GO Anno	$\mathcal{K}_{SIM}^{ext}, \mathcal{K}_{FUN}^{ext}$	$\mathcal{K}_{SIM}^{int}, \mathcal{K}_{FUN}^{int}$	Geometric Consistency
GO-CAN-MAH	GO Anno, CAN Anno	$\mathcal{K}_{SIM}^{ext}, \mathcal{K}_{FUN}^{ext}$	$\mathcal{K}_{SIM}^{int}, \mathcal{K}_{FUN}^{int}$	Geometric Consistency
GO-REL-PROP	GO Anno, CAN Anno	$\mathcal{K}_{CON}^{ext}, \mathcal{K}_{FUN}^{ext}$	$\mathcal{K}_{CON}^{int}, \mathcal{K}_{FUN}^{int}$	Relevance Propagation
Leukemia-FILT	CAN Anno	\mathcal{K}_{FUN}^{ext}	\mathcal{K}_{GEO}^{int}	Geometric Consistency

METHODS	DESCRIPTION
SPEC	Expression of genes are used to construct \mathcal{K}_{GEO}^{int} with Mahalanobis distance.
Fisher Score	\mathcal{K}_{CAT}^{int} (Label information), is used with Fisher Score to select genes.
Pathway-FILT	Genes in selected pathways are used to filter the data, \mathcal{K}_{FUN}^{int} is obtained on the filtered data.
GO-REL-VOTE	GO terms are weighed according to their relevance, they used to rank genes. See Section 4.1.3
GO-MAH	GO based gene similarity is used to construct Mahalanobis distance to extract \mathcal{K}_{GEO}^{int} .
GO-CAN-MAH	Similar to GO-MAH, but only cancer related GO terms are used to calculate gene similarity.
GO-REL-PROP	Relevance is propagated on the graph constructed from the GO based gene similarity. See Section 4.1.2.
Leukemia-FILT	Use genes with ALL-related functions to filter the data, and \mathcal{K}_{GEO}^{int} is obtained on the filtered data.

*C. D. Pitta, et al. A leukemia-enriched cDNA microarray platform identifies new transcripts with relevance to the biology of pediatric acute lymphoblastic leukemia. *Haematologica*, 90:890–898, 2005.

PERFORMANCE COMPARISON

Performance comparison for gene ranking methods using one type of knowledge with using multiple types of knowledge.

METHODS	ACC-10	ACC-30	ACC-50	ACC Ave	Sihanno	HIT _{case}	HIT _{test}	REL _{pos}
SPEC	0.64	0.66	0.83	0.65	797	2	0	21
Fisher Score	0.97	0.97	0.97	0.97	823	8	2	14
Pathway-FILT	0.61	0.81	0.89	0.81	807	4	0	19
GO-REL-VOTE	0.56	0.69	0.83	0.64	7686	26	8	25
GO-MAH	0.69	0.80	0.86	0.82	759	3	0	14
GO-CAN-MAH	0.62	0.83	0.86	0.80	2096	5	1	17
GO-REL-PROP	0.70	0.78	0.86	0.74	7688	22	15	33
Leukemia-FILT	0.55	0.62	0.64	0.62	687	4	1	20
KOGS _{meta}	0.91	0.97	0.97	0.96	1723	6	2	16
KOGS _{prob}	0.97	0.94	0.94	0.95	6954	21	12	20
KOGS _{prob-sup}	0.94	0.91	0.91	0.93	7766	24	17	29

FURTHER STUDY ON BIOLOGICAL RELEVANCE

The biologically relevant genes in the top 50 gene list provided by KOGS_{Prob-SUP}. The upper part of the table contains the genes whose relevance to leukemia has been confirmed by the literature. And the lower part of the table contains the genes, whose relevance is unknown but cannot be ruled out.

LMO1, CBFA2T3, TYROBP, STAT5B, IGFBP3, JUN, USP33, GSN, BTG1, TFRC, PTK2, PDE7A, TIMP1, AKT1, FLT1, CEPBD, TIMP2, TIMP4, TYK2, CDK4, SERPIN2, PRKACA, NCOR1, SIVA1, BRD8, CAPN7, SPATA2, PRKAR1A, PPARA

The results show that the framework can select genes bearing both statistical significance and biological relevance.

6. Conclusion

In this work, we proposed KOGS, a general framework for knowledge oriented gene selection to convert different types of external knowledge to internal knowledge for ranking genes. Given multiple gene ranking lists, KOGS can aggregate them to form a final list considering various gene relevance. Experimental results demonstrated the methods derived from KOGS can select genes bearing both statistical significance and biological relevance.

7. Acknowledgment

This work is in part sponsored by NSF-0812551 to HL.